computation will provide significant economies of scale and the requisite fidelity and quality. A centralized, computationally integrated facility also will maximize efficient use of resources.

Production Capability

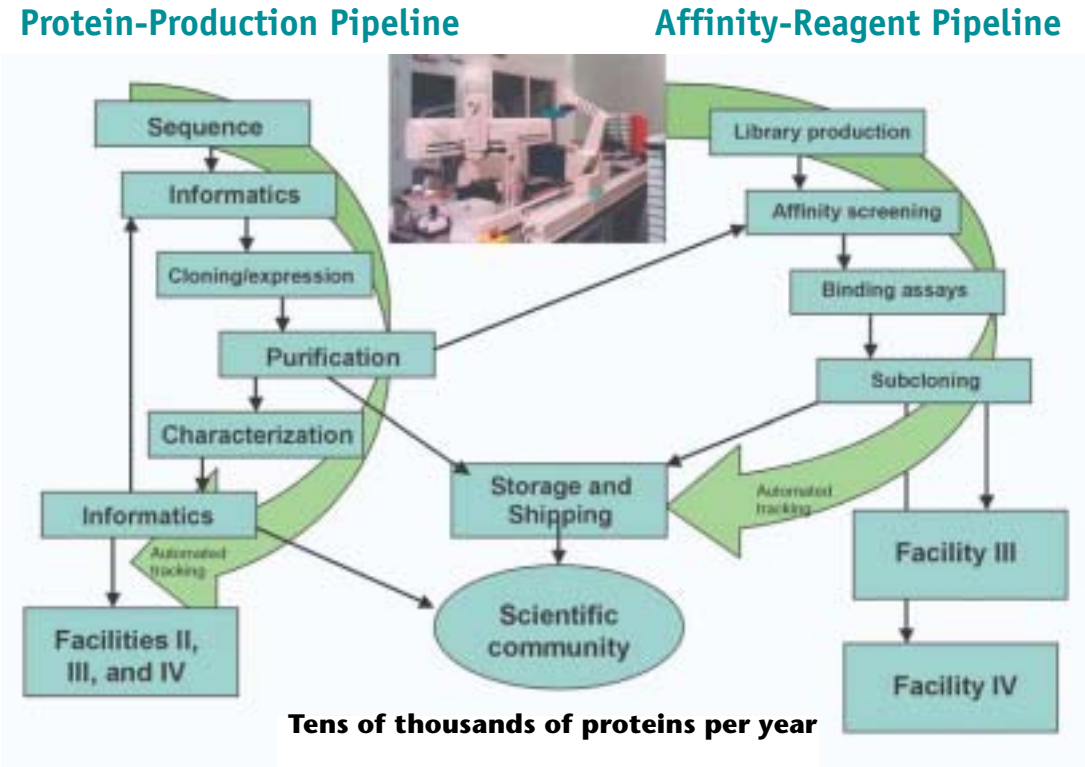
Scale

- 10,000 to 25,000 purified proteins per year
- Milligram quantities of each product
- Soluble, full length, and natively folded
- High rate of success (>95%) for production of proteins

Scope

- Proteins from genetic sequences of target organisms
- Protein variants
  - Isotopically labeled proteins
  - Post-translationally modified proteins
  - Proteins with unknown cofactors
  - Proteins incorporating nonstandard amino acids
  - Site-specific mutant arrays (high-throughput mutagenesis)
  - Fusion tag arrays
  - Affinity reagents (e.g., antibody domains) for every protein produced

Facility I: Production and Characterization of Proteins



GTL Facility I will use highly automated processes to mass-produce and characterize proteins directly from genome information and affinity reagents ("tags") to identify, track, quantify, manipulate, capture, and monitor the proteins. Protein production (left grouping) requires comparative, informatics-guided selection of appropriate targets, followed by cloning their genes, inducing expression in cell-free extracts or cells, and purification. Systematic biophysical characterizations will be completed on each protein. Once available, proteins will be used to generate specific affinity reagents (right). Current approaches rely on selecting affinity reagents from a diverse library and subsequent amplification. Expression clones, proteins, affinity reagents, and characterization data will be used by Facilities II, III, and IV to study molecular machines by mass spectrometry, imaging, and modeling approaches.

The currently funded NIH distributed structural genomics centers have had some success in producing proteins, and these results will be useful in developing Facility I. The focus of NIH efforts to exclude characterization beyond 3D structures means that many of the proteins selected for expression are either small or not full length and poorly represent the many types and classes of proteins. Key classes, comprising a third or more of the total, are significantly disordered in solution and thus are neither amenable to, nor meaningfully characterized by, conventional structural determinations. Yet, disorder in proteins is emerging as an increasingly important factor in determining function—particularly in the assembly of protein partners into molecular machines. This key process very often is mediated by disorder-to-order transitions at the binding interfaces. Facility I will provide general biophysical characterizations of full-length proteins that will, among other things, allow their general structure (whether ordered or disordered) to be defined. Integration of data obtained in Facility I, with appropriate informatics efforts, eventually will allow protein disorder to become a useful tool to predict binding partners and aspects of protein function.

The production of multiple high-affinity, high-specificity affinity-tag reagents for each protein presents its own enormous challenges. Several promising approaches to this problem are under development worldwide, although none have yet emerged as economical and reliable solutions to the high-throughput needs of GTL. Overcoming this obstacle is therefore a major target for GTL pilot studies and for this facility in particular. Computational tools will be employed to provide an initial understanding of the genes and protein complement of each microbe studied and to estimate protein function and organism capabilities based on the catalog of previously analyzed microbes. This analysis will identify novel proteins and those involved in previously characterized molecular machines. Prior knowledge and curated and archival data will be used to build predictive models of protein function and behavior. These data and models will be readily accessible to enable studies of proteins and protein complexes.

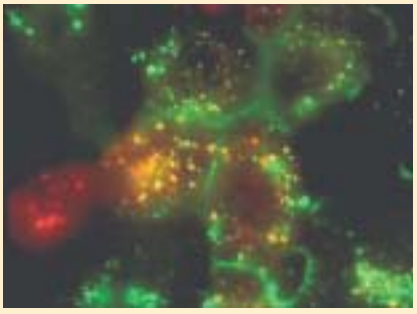
Proteins, tags, and data produced in Facility I are needed by Facilities II, III, and IV to capture the molecular machines for mass spectrometry (MS) analysis and to identify the machines' components. They also are needed for cellular-imaging studies, reconstitution of molecular machines, and verification of models. Characterization data will provide vital insights about which affinity tags are likely to disrupt or not disrupt cellular protein function.

Key Technologies Needed

- Capabilities for large-scale production and purification of milligram quantities of active, full-length proteins. This task includes difficult proteins to produce and characterize, such as those that are unstable or disordered, particularly in the absence of their binding partners.
- Reliable and high-throughput methods to successfully refold proteins.
- High-throughput methods to characterize proteins by multiple, independent biophysical assays under several standardized conditions.
- High-throughput methods to produce and characterize multiple affinity tags for each protein.
- Laboratory Information Management System (LIMS) for tracking and managing samples, tags, and production conditions, as well as sample export to the research community.
- Computational tools for efficiently collecting, analyzing, and interpreting the above-noted production and characterization data. These will

Lighting Up Proteins in Cells

Fluorescent labeling using tags such as those produced in Facility I provides a way to study the functions of specific proteins in living cells. It allows direct observation in real time of the protein's location in time and space as well as the interactions of proteins with each other and with other cellular elements. The thousands of proteins produced by automated methods in Facility I will provide scientists a means to study these components in living cells, both in GTL user facilities and in their own labs. This image shows the distribution and expression levels of the two proteins in a grouping of cells. The yellow emissions indicate simultaneous green and red emissions and thus reveal regions where both proteins are present (within the microscope's resolution of ~0.4 μm).



include tools for analyzing successful and unsuccessful expression; generating initial comparative genome analysis to determine genes, proteins, and regulatory elements; and identifying previously known protein associations with other proteins and ligands.

- Data-management systems for capturing knowledge about previously examined microbes; detailed protein, machine, and regulatory-element comparisons; and protein-production data and conditions.

Project Description

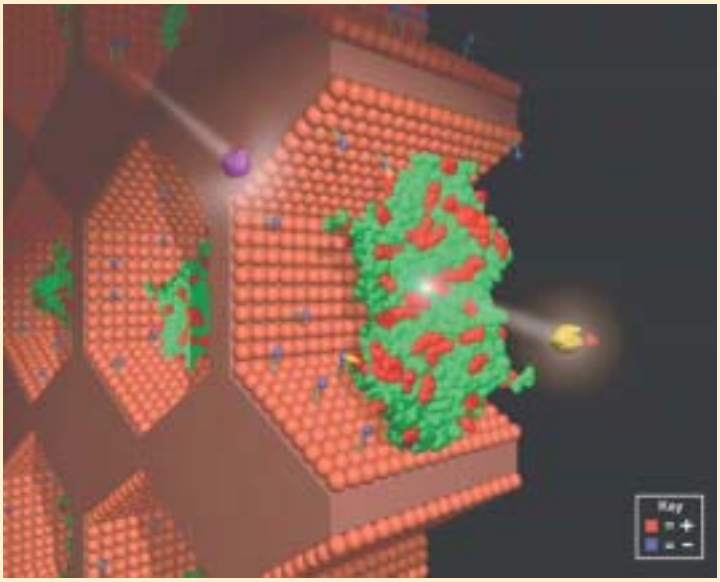
Facility I will revolutionize how proteins, affinity reagents, and associated characterization data become available to the scientific community. A sophisticated informatics capability tracking all aspects of production and characterization means that crucial data and reagents could be applied to a myriad of scientific problems. This computational infrastructure will enable use of the DNA sequence to predict the following for each protein: efficient and successful production methods, likely binding partners, and ultimately information about the functions of each gene. Achieving this goal will require experience and the data created from production and characterization of tens of thousands of proteins. Automation and computationally based insights are keys to achieving high throughput at steadily declining costs, just as they were in DNA sequencing. Proteins are more difficult to handle than DNA, so no single production method and characterization scheme will be applicable to all proteins. Thus, several methods will be developed simultaneously.

Whichever method is selected, nearly all protein production is based on transcription from DNA. This DNA is produced via cloning or possibly direct chemical synthesis of the gene encoding the desired protein. Facility I, as part of its function as a national resource, will develop a sequence-verified library of publicly available protein-coding microbial genes. This library would be available for translation into protein or usage in transformational studies by the other facilities or the larger scientific community.

A Possible Application of Proteins Produced in GTL Facilities

Harnessing Enzymes to Inactivate Contaminants and Generate Energy

Miniaturization technologies can be combined with biological components such as enzymes to create novel systems that use an array of microbial processes but do not require living cells. In this figure, the enzyme organophosphorus hydrolase (OPH) is embedded in a synthetic nanomembrane (mesoporous silica) that enhances its activity and stability [*J. Am. Chem. Soc.* **124, 11242–43 (2002)]. The pores (up to 30 nm) were functionalized with a few negative charges to aid in holding the positively charged OPH in place. The figure shows the enzymes, pores, functionalization (blue balls), and contact between negatively charged pores and some positively charged regions of the enzyme (glowing blue and red balls).**



A unique feature of OPH is its ability to inactivate an unusually broad range of chemicals (substrates), including several nerve gases (e.g., sarin) and pesticides. In addition to using individual enzymes, the technology also may be useful for immobilizing clusters of enzymes in particular metabolic pathways. Application such as this could enable development of efficient enzyme-based ways to produce energy, remove or inactivate contaminants, and sequester carbon to mitigate global climate change. It also could be highly useful in food processing, pharmaceuticals, separations, and the production of industrial chemicals.

High-throughput protein production requires multiple approaches based on cell-free and cellular methods. Direct chemical synthesis may someday represent a viable alternative, although refolding into active protein remains a major unsolved problem.

Cell-Free Systems

Cell-free expression systems, such as those based on wheat germ extracts or *Escherichia coli,* hold the greatest potential for full automation and hence lower costs and high throughput. Successful efforts in Japan using these extracts have yielded thousands of proteins per year. The ability to automate these systems and the potential to incorporate labeled or nonstandard amino acids warrant a substantial investment in these highly promising and flexible in vitro methods in Facility I.

Cell-Based Expression Systems

Large-scale cell-based expression systems have been used with some success worldwide in structural genomics centers and elsewhere. These approaches cannot, however, be as readily automated as cell-free systems. They also suffer from fairly low success rates. Partly for this reason, yeast and other eukaryotic expression systems have been developed and are typically resorted to for proteins that fail in *E. coli* –based systems.

Chemical Synthesis

Solid-state chemical synthesis is a possible approach for important proteins that fail in all DNA-based expression systems. Currently, this method can produce peptides up to 50 amino acids in length, but longer peptides are made at ever-diminishing efficiencies. Full-length proteins might be synthesized through chemical ligation of multiple peptides. This is, however, a costly procedure, and refolding into active protein remains a major unsolved problem. This technique has the advantage of producing milligrams of proteins labeled by incorporation of isotopes, chemical modifications, unnatural amino acids, or other chemical groups. Facility I will implement this approach in later years.

Protein Purification

Protein purification (*after expression*) presents a number of challenges, particularly in a high-throughput environment. In Facility I, substantial reliance will be placed on experience-based informatics methods to guide the purification strategy for each protein—with the expectation of achieving significant improvement as the database expands. Automated protocols aimed at eliminating centrifugation will be developed since this step accounts for the major bottleneck in current protein-production protocols.

Characterization of Proteins

A key and largely unique goal of Facility I is stabilization and extensive characterization of each produced protein under well-defined conditions. Given the investment in each expressed protein and its scientific value, subjecting each to a substantial suite of assays is planned: solubility measurements by light scattering; probing conformation and disorder by circular dichroism; HSQC nuclear magnetic resonance; partial proteolysis and isotope exchange coupled with MS; small-angle scattering, dye-binding and spectrofluorimetry; surface plasmon resonance; calorimetry; and protein chip–based binding measurements to extracts, proteins, ligands, and nucleic acids. This and the large number of assays contemplated and the high targeted rate of protein production imply the need for a careful and systematic development of high-throughput, robotic approaches. Facility I will be responsible for characterization data that is scientifically useful.

Affinity-Tag Production

Once proteins become readily available, they enable production of protein-specific affinity reagents. Several different and complementary approaches to generating affinity reagents are under development worldwide. These include phage and yeast display systems and aptamers. All have promising features, but none of these technologies has been developed sufficiently to satisfy the high-throughput requirements necessary for this facility. Further developmental areas include improved reagent stability and specificity; improved multiplex screening protocols; and rapid, high-throughput affinity-maturation techniques. The reagents also will be evaluated to determine where they bind to their protein targets and whether they disrupt the target's function, thereby dictating how different affinity reagents can be used. Development of modular affinity reagents also would be extremely useful; selected binding domains could be inserted into standardized structural modules to allow affinity reagents to be generated rapidly for different purposes such as protein isolation or live-cell imaging.

The most useful affinity reagents probably will be proteins themselves. They can be produced and characterized using technologies already developed for bacterial proteins. Because they will be standardized reagents, however, processes can be developed to allow for their rapid and large-scale production, enabling their distribution to scientists worldwide and greatly enhancing the scientific impact of reagents generated in the facility.

Computational Infrastructure

Central to this facility will be computational resources that provide an initial estimate of genes and proteins in a genome and identify proteins with known function or associations in previously characterized molecular machines. Computational tools will be utilized to estimate the capabilities of each new genome and help prioritize proteins for production. Systems are needed that will allow tracking of samples from the DNA constructs to the incorporation of data into gene-annotation databases. LIMS will be used to track samples (via bar coding) and incorporate all information relevant to sample history. A suite of computational tools for automated analysis and archiving of protein production and characterizaton data will be established to feed bioinformatics tools that will interpret these data.

Impacts on Science and DOE Missions

The postgenomic era of biology will focus on the genome and how it creates and utilizes proteins. Availability of thousands of microbial proteins and specific, high-affinity tags will substantially free the nation's scientists to apply their energies to understanding how microbial proteins function. Microbial capabilities then can be harnessed for exciting applications to DOE missions. Examples are hydrogen fuel cells based on hydrogenase enzymes supplied and maintained by clusters of associated molecular machines or environmental cleanup based on enzyme inactivation of toxic chemicals. This facility will provide unique resources and technologies important to systems biology.

Probabilities for Success

- DOE has the capabilities to establish a centralized facility that combines biological, physical, and computational sciences at the scale required for successful production and characterization of vital proteins and tags.
- DOE has supported many of the necessary components as pilot programs including the following:
  - Pilot facilities for protein production have been developed in conjunction with structural genomics research at several national laboratories including Argonne, Brookhaven, Los Alamos, Pacific Northwest, and Berkeley and are supported in partnership with the NIGMS Protein Structure Initiative. These facilities, which produce milligram quantities of hundreds of proteins per year, are developing the technologies needed to bring protein production to the next level of automation, completeness, and reproducibility.
  - The production of affinity reagents for proteins is being explored at several national laboratories, including Pacific Northwest, Argonne, and Los Alamos. These reagents cannot be designed but rather are selected from very large combinatorial libraries of reagents. The national laboratories are providing the necessary experience for developing novel libraries and automating their screening for the high-throughput production of affinity tags.
- Some aspects of protein production by cell-free systems can be modeled on facilities in Japan that are willing to assist DOE in establishing its significantly larger and more comprehensive Facility I.

Facility II: Whole Proteome Analysis

All the proteins encoded in the genome make up an organism's "proteome." The cell does not generate all these proteins at once but rather the set required at a particular time to produce the functionality dictated by environmental cues and the organism's life strategy. To make use of any microbe's capabilities, we must understand the principles of these processes.

GTL Facility II will characterize the expressed proteomes of diverse microbes under different environmental conditions as an essential step toward determining the functions and interactions of individual proteins and sets of proteins.

Strategic Intent

While the information content of the genome is relatively static, the processes by which families of proteins are produced and molecular machines are assembled for specific purposes are amazingly dynamic, intricate, and adaptive. Microbes deploy an interacting and changing panoply of proteins to carry out the myriad processes necessary for life. In addition to the networks and structures that do the cell's core work, numerous other machines and networks serve as sensors and regulators for the production and control of all the cell's elements. The principles for deployment of cellular capabilities can be deduced much more readily if a cell's protein makeup can be measured and correlated to external stimuli and to cellular responses. Thus, characterizing a microbe's expressed collection of proteins is an important first step toward deciphering the principles by which the genome regulates the assembly and functioning of molecular machines. However, a microbe typically expresses thousands of distinct proteins at a time, and the abundance of individual proteins may differ by a factor of a million. Technologies only recently have emerged that can successfully measure this breadth and dynamic range, and a substantial technical and computational infrastructure is required for their use (see "*Deinococcus*" sidebar, p. 18, on the FTICR Mass Spectrometer and Accurate Mass Tags results].

The Whole Proteome Analysis Facility (Facility II) will employ state-of-the-art techniques to generate data with high efficiency and validity; it will bring the same economy of scale to proteomics that centralized centers brought to gene sequencing. This facility will build on information obtained from the genome sequence to identify, in a snapshot fashion, the thousands of dynamically changing proteins expressed in a living microbe.

Project Purpose and Justification

The new era of systems biology requires the generation of massive amounts of whole-systems data collected under highly controlled and reproducible conditions from the point of bacterial cell growth to data archiving and dissemination. No facilities currently exist with the range and scale of capabilities and capacities required to collect these types of data. Further, centralizing the analysis of proteins within a specialized facility in a manner analogous to today's genome-sequencing centers would allow us to conduct these assays with higher efficiency, fidelity, and throughput than could be accomplished in the laboratories of individual investigators.

This facility will establish capabilities that will permit the annual measurement of the expression levels of proteins in bacteria grown under hundreds of different conditions. This will make possible the integrated study of thousands of proteins of numerous microbial species under a significant range of environmental changes. For the first time, we will be able to visualize a cell's systems-level response to these changes and identify the critical molecular changes that result from those conditions.

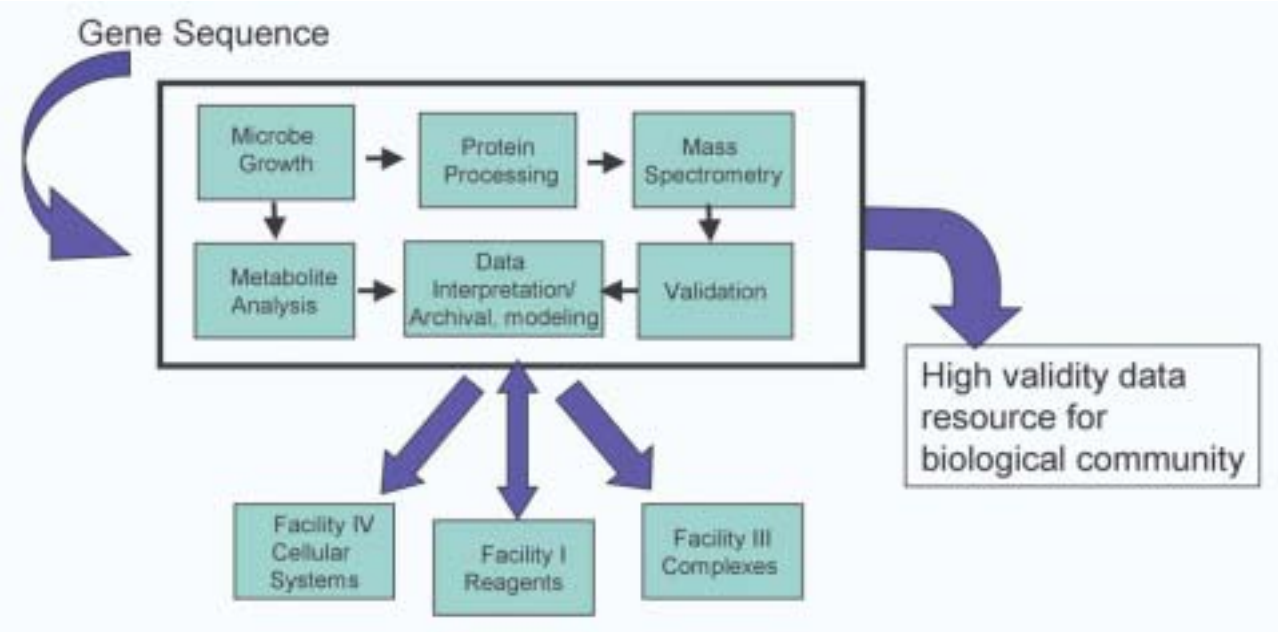
Many of the technologies needed for whole-proteome analysis have been successfully demonstrated in pilot projects funded by the BER program. Facility II will incorporate those technologies and others within suites of integrated analytical and computational tools. This facility will include capabilities to grow microorganisms under controlled conditions, isolate proteins from cells, and identify and quantify proteins using MS and other analytical techniques. Computational tools will be employed to interpret and archive data and to build predictive models of subsystems that control protein expression and affect cellular response to conditions. Computational modeling and simulation will be used interactively with experimental data collection to achieve a quantitative understanding of the components and parameters affecting expression. This comprehensive knowledge—cap-

tured in data, models, and simulation codes—will be disseminated to the greater biological community to enable studies of microbial systems biology.

Key Technologies Needed

- High-capacity cultivation systems, including controlled and automated chemostats for growing microbes under defined conditions in batch or continuous culture and capabilities to handle difficult-to-cultivate microbes.
- High-throughput techniques for preparing microbial samples before proteome analysis that incorporate integrated sample-processing systems, robotics, and automation.
- Capabilities for large-scale analysis of microbial proteomes, incorporating demonstrated MS-based methods to produce high-quality data. Analytical and computational methods for detecting and quantifying proteins modified by such methods as phosphorylation and methylation.
- High-sensitivity analytical tools for high-throughput analysis of metabolites, lipids, carbohydrates, and other cellular constituents.
- High-performance computational tools and codes for efficiently collecting, analyzing, and interpreting whole-proteome data. Tool capabilities, including data clustering, expression analysis, and genome annotation, would be closely linked to the advances in computing infrastructure being proposed by DOE.
- Computational tools for abstracting network and pathway information from expression data and genome annotation and for building mathematical models that represent subcellular systems responsible for protein expression and proteome state (including modified proteins) as a function of conditions. Simulation would be used to evaluate the state of knowledge contained in these models and validate the accuracy of experimental parameters.
- Data-management systems for archiving large amounts of data that may exceed petabytes and that can be accessed easily by a large community of users. Databases of expression measurements, metabolome measurement, and networks and pathway systems, models, and simulation codes

Facility II: Whole Proteome Analysis



GTL Facility II will characterize the expressed proteomes of diverse microbes under different environmental conditions as an essential step toward determining the functions and interactions of individual proteins and sets of proteins. Proteins identified in Facility II will help guide the production of reagents in Facility I. Reagents from Facility I also will be employed to assist in verifying the identities of proteins studied in Facility II. Proteomics data from Facility II will enable the identification of protein complexes in Facility III. Similarly, data from this facility will enable the study of proteins, metabolites, and other cellular constituents in the cellular systems studied in Facility IV.

would be developed.

- Research activities to develop new capabilities to improve the throughput, sensitivity, and information content of analytical tools and research in complementary computational methods to better interpret and visualize the results of MS measurements and other types of complex experimental data.

Project Description

This facility will consist of specially designed laboratories to house resources for cell growth; high-throughput sample preparation; state-of-the-art analytical instrumentation, including a suite of mass spectrometers; and computational infrastructure for sample management, data analysis (leveraging DOE's high-performance computing infrastructure), curation, archiving, and dissemination.

Proteome-Expression Systems

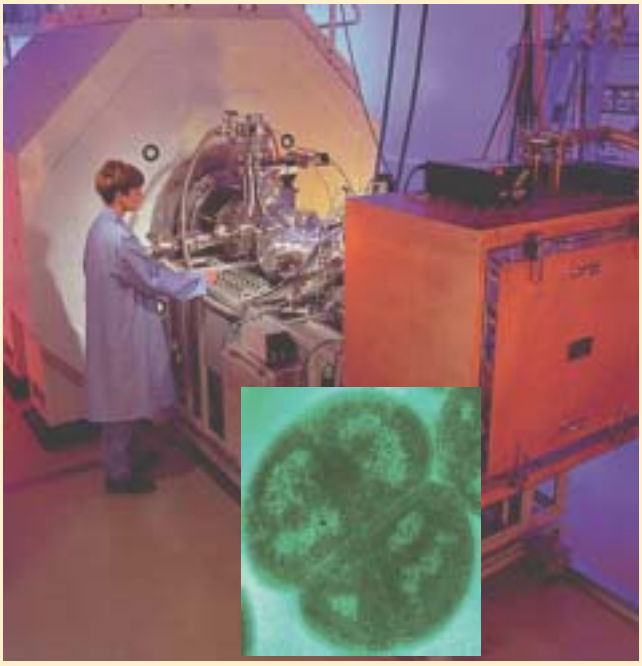
Cells will be grown under controlled conditions to produce proteins in sufficient quantities for analysis and would require fermentation technologies. The goal will be to rapidly assay proteomes from the target organism grown under a range of well-defined conditions. These multiple expressed proteomes will aid in rapid hyperannotation of new genomes by directly linking each genome to the expressed

A Possible Application of Knowledge Gained from GTL Facilities

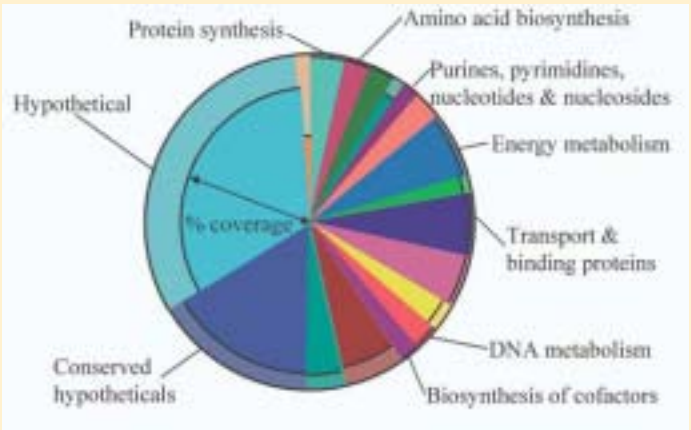
Deinococcus radiodurans: *Offering Promising Solution to Site Cleanup*

Dubbed the "world's toughest bacterium" by the *Guinness Book of World Records*, *D. radiodurans* (inset photo) can withstand extremely high levels of radiation and has excellent potential for use in stabilizing radioactive contaminants. Understanding the proteins involved in these capabilities may one day enable use of this microbe or its components in environmental cleanup. A BER pilot project to determine the proteome (the complement of proteins) of this microbe generated the most complete coverage obtained yet for any organism (*PNAS*, Aug. 20, 2002). Some 1900 proteins were identified, representing 83% of proteins predicted from the genome sequence (see pie chart below). The Office of Science's Microbial Genome Program provided the genomic-sequence information.

This unprecedented coverage was achieved using a new high-throughput mass spectrometer based on Fourier-transform ion cyclotron resonance (FTICR) developed at Pacific Northwest National Laboratory (photo). The system relies on a two-step process that first uses tandem mass spectrometry to identify biomarkers (accurate mass tags or AMTs) for each protein (see pie chart below). With this method, thousands of proteins were identified in a matter of hours. Identifying AMTs allows changes in the proteome to be monitored extremely efficiently.



Functional Classifications of Proteins



proteome under multiple growth and stress conditions. These whole-cell samples will be archived for subsequent analysis. Associated with these expression laboratories will be a range of analytical equipment, including chromatographs and mass spectrometers, to analyze other cellular components, such as metabolites. Significant investment will be also made in developing methods for working with microbes.

Protein-Sample Processing

Highly automated processes will be established to perform initial isolations of proteins from microbes, final sample preparation (e.g., desalting, buffer exchange, and sample concentration), and enzymatic digestion of samples as required for analysis. Sample-handling steps will be minimized using robotic and liquid-handling systems to eliminate many of the current bottlenecks in preparing samples for analysis.

Mass Spectrometry of Proteins

MS will be used to measure the molecular masses and quantify both the intact proteins and the peptides produced by enzymatic digestion of the proteins. Identification of the expressed proteins will require both moderate-resolution "workhorse" instruments, such as quadrupole ion traps, and high-performance mass spectrometers capable of high mass accuracy; the latter includes Fourier transform ion cyclotron resonance (FTICR) mass spectrometers or orthogonal injection time-of-flight mass spectrometers (called Q-TOF). These instruments will be interfaced with liquid chromatographs equipped with autosamplers to permit online separation of components prior to MS analysis. The data output of these instruments will require extensive dedicated computational resources for data collection, storage, interpretation, and analysis. [see "*Deinococcus*" sidebar on FTICR MS, p. 18]

Quality Assurance and Other Analytical Techniques

Quality assurance will be an important component within this facility. Validation of the identities and quantities of proteins present would be performed using complementary analytical and radiolabeling techniques and would require such equipment as high-performance liquid chromatography, spectrophotometers, gel electrophoresis, mass spectrometers, expression arrays, imaging tools, and computer workstations.

Computational Resources and Capabilities

Central to this facility will be an array of computational resources that will be employed to track samples and handle all aspects of data collection, storage, interpretation, and analysis. Databases and tools will be established for use by the biological community to access the data and models produced by the facility. LIMS will be used to track samples and incorporate all information relevant to sample history. A suite of computational tools for automated analysis and archiving of mass-spectral (protein-expression) data will be developed to feed bioinformatics tools that will interpret these data and identify proteins and their post-translational modifications. Such computational resources also will use data input from analyses established for quality assurance, including data from multiple MS runs, gel electrophoresis, and other assays.

This facility will develop and deploy large-scale data-analysis tools and infrastructure, tools for modeling and simulating protein expression based on collected MS data, petabyte-data management, and user interfaces for dissemination of data and proteome models to the community. For large-scale data analysis, modeling, and simulation, the facility will employ DOE's high-performance computing infrastructure. Additional data generated by other facilities also will be incorporated in these databases to enhance understanding of protein function in cells.

Optical Analysis and Cell Sorting

Biological systems are inherently inhomogeneous; measurements of the average proteome's expression profile for a collection of cells cannot be related with certainty to the protein-expression profile of any particular cell. This is especially true for proteins found in small amounts that may be expressed either at low levels in most cells or at high levels in only a small fraction of the cells. For these reasons, Facility II will use fluorescent probes developed in conjunction with Facility I to enable quantitative imaging of proteins within living cells. To achieve the most direct correspondence between imaging data and proteomic data, Facility II will conduct some of these measurements on the actual cultures used for the MS analyses. As a refinement, flow cytometric techniques will be used to separate various cell states to allow specific groups of cells to be studied from heterogeneous cultures. Other studies can be conducted using imaging capabilities in the other GTL facilities and at universities and national laboratories.

Technology Development

An important part of this facility is the development of new biological, analytical, and computational tools to improve the sample throughput and information content required for the GTL program. Although current state-of-the-art techniques allow the analysis of many proteins within a cell, new approaches will be required to permit efficient analysis of the full range of proteins without special handling. These approaches would include methods for isolating and analyzing membrane proteins, improving methods for quantification, and developing technologies for analyzing proteins from a single cell. Few computational tools are available to analyze these types of data and translate them into functional information and robust models of cellular subsystems. Many research activities need to be conducted in close association with this facility to meet the needs of the GTL program; numerous innovations, however, will come from individual investigators across the scientific community.

Impacts on Science and DOE Missions

Many microbial processes may be useful for DOE missions such as energy production, carbon sequestration, and environmental cleanup. Relating the microbe's genome to the specific functions conducted by its proteins may make possible the activation or deactivation of particular processes. This and other capabilities will enable us to optimize microbial systems or learn how to construct nonliving systems that mimic processes found in microbes for specific applications—a precursor to harnessing such processes to meet the needs of DOE's missions.

Understanding changes induced in a microbe's protein-expression profile by different environmental conditions will serve as a basis for identifying the function of individual proteins. This will provide the first step toward understanding the function of the complex network of processes conducted by a microbe.

Data from Facility II will be used to develop models for predicting microbial responses to different environments and for using these capabilities in practical applications that meet DOE mission needs. Further, if we can understand the processes in relatively simple microbial cells and extend this understanding to communities of microbes, then we can apply this knowledge to higher organisms.

Probabilities for Success

- Experience in the Human Genome Project taught us that biological analyses, like DNA sequencing, can be standardized, automated, and made high throughput and cost-effective. Similarly, large-scale protein analysis can provide high-quality data to the general scientific community.
- Research under way in the Genomes to Life program, industry, and across the federal government will identify many individual technology components and methods needed to make large-scale proteome analysis practicable.
- A BER-supported pilot research project already has characterized more than 80% of the computationally predicted proteome of one microbe under a variety of experimental conditions. Similar work is under way on *Shewanella*, *Rhodopseudomonas*, and *Prochlorococcus*.

As this facility progresses, multidisciplinary teams of physical, biological, and computational scientists will continuously expand its capabilities to improve the throughput and information content of analytical data and provide new computational models for predicting protein expression in microbial cells.